AMENDMENTS TO THE CLAIMS:

This listing of claims will replace all prior versions, and listings, of claims in the application:

LISTING OF CLAIMS

1. (Currently Amended) A method, in a network comprising a primary server and a plurality of offload servers, for dynamic offloading of processing requests from said primary server to any one of said plurality of offload servers, the method comprising ~~the~~ steps of:

determining a load on said primary server;

if the load on said primary server is less than a first threshold, serving processing requests at said primary server;

only if the load on said primary server exceeds said first threshold, then offloading at least a portion of said processing requests to any one of said plurality of offload servers while said primary server continues to serve a remainder of said processing requests, wherein ~~each of~~ said any one of said plurality of offload servers ~~is configured to process~~ processes said at least a portion of said processing requests and is associated with ~~a respective~~ an offload threshold and the at least a portion of said processing requests is the only work handled by said any one of said plurality of offload servers, wherein the offloading is performed in accordance with a ~~said~~ respective offload threshold for each of the plurality of offload servers, such that if said respective offload threshold is exceeded for every one of the plurality of offload servers, said offloading is stopped until a load on one of said plurality of offload servers falls below said respective offload threshold; and

if the load on said primary server exceeds a second threshold, throttling at least one of said processing requests,

wherein serving the processing requests at said primary server includes returning a page to a user wherein all embedded objects in the page have links to said primary server; and

wherein offloading at least a portion of the processing requests to any one of said plurality of offload servers includes serving a base page at said primary server in which links for embedded objects point to any one of said plurality of offload servers.

2. (Original) The method of claim 1 wherein said load comprises bandwidth utilization and said first threshold is a network bandwidth utilization of said primary server.

3. (Previously Presented) The method of claim 1 wherein said load comprises CPU utilization and said first threshold is a central processing unit (CPU) utilization of said primary server.

4. (Cancelled)

5. (Previously Presented) The method of claim 1 wherein offloading at least a portion of the processing requests to any one of said plurality of offload servers includes routing an incoming Web request to a selected offload server.

6. (Cancelled)

7. (Previously Presented) The method of claim 1 wherein throttling at least one of said processing requests includes returning a page to a user indicating that a server is overloaded.

8. (Previously Presented) The method of claim 1 wherein throttling at least one of said processing requests includes dropping the at least one of said processing requests without returning any information to a user.

9. (Previously Presented) The method of claim 1 wherein throttling at least one of said processing requests includes returning a page to a user indicating that a server is

overloaded if said load exceeds said second threshold, and dropping said at least one of said processing requests if said load exceeds a third threshold.

10. (Previously Presented) The method of claim 1 wherein a determination of which of said plurality of offload servers that at least a portion of said processing requests is to be offloaded to is based on one or more of a group including; a client identity, a client gateway (Internet Protocol) address, a price of offload service, or a current or previous load on the any one of said plurality of offload servers.

11. – 31. (Cancelled)

32. (Currently Amended) A method for allocating processing requirements on an Internet Protocol network between a primary server and a plurality of offload servers, comprising:

periodically evaluating processing requests to determine a load on said primary server;

if said load exceeds a first threshold, for a predetermined period of time directing at least one of said processing requests to any one of said plurality of offload servers while said primary server continues to serve a remainder of said processing requests, wherein ~~each~~ said any one of said plurality of offload servers ~~is configured to process~~ processes said at least one of said processing requests and is associated with ~~a respective~~ an offload threshold and said at least one of said processing requests is the only work handled by said any one of said plurality of offload servers, wherein the directing is performed in accordance ~~said~~ with a respective offload threshold for each of the plurality of offload servers, such that if said respective offload threshold is exceeded for every one of the plurality of offload servers, said offloading is stopped until a load on one of said plurality of offload servers falls below said respective offload threshold;

only if said load does not exceed said first threshold, directing said processing requests to said primary server; and

if the load on said primary server exceeds a second threshold, throttling at least

one of said processing requests, wherein directing said processing requests to said primary server further includes returning a page to a user wherein all embedded objects in the page have links to said primary server; and

directing at least one processing request to any one of said plurality of offload servers further includes serving a base page at said primary server in which links for embedded objects point to said any one of said plurality of offload servers.

33. (Previously Presented) The method of claim 32 wherein said load comprises network bandwidth and said first threshold is a measure of network bandwidth utilization of said primary server.

34. (Previously Presented) The method of claim 32 wherein said load comprises central processing unit (CPU) utilization and said first threshold is a measure of CPU utilization of said primary server.

35. (Cancelled)

36. (Previously Presented) The method of claim 32 wherein directing at least one processing request to any one of said plurality of offload servers further includes routing an incoming Web request to a selected offload server.

37. (Previously Presented) The method of claim 32 wherein said throttling at least one of said processing requests comprises returning a page to a user indicating that a server is overloaded.

38. (Previously Presented) The method of claim 32 wherein said throttling of at least one of said processing requests comprises dropping the at least one of said processing requests without returning any information to a user.

39. (Previously Presented) The method of claim 32 wherein the throttling of at least

one of said processing requests comprises returning a page to a user indicating that the primary server is overloaded if the load exceeds the second threshold, and further comprising dropping the at least one of said processing requests if the load exceeds a third threshold.

40. (Previously Presented) The method of claim 32 further including determining which of said plurality of offload servers said at least one of said processing requests is to be offloaded to based on one or more of a group including: a client identity, a client gateway (Internet Protocol) address, a price of offload service, or a current or previous load on the any one of said plurality of offload servers.

41. – 42. (Cancelled)